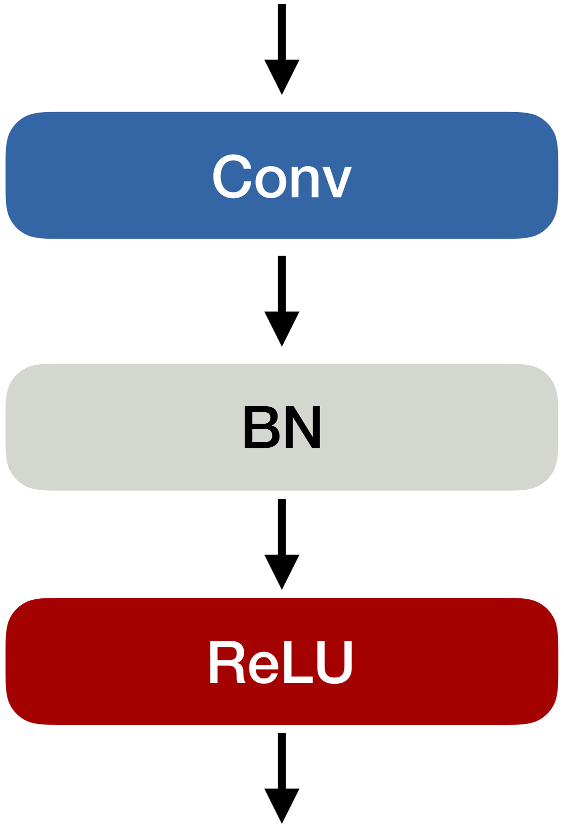
Batch normalization

Batch normalization

- Make activations zero mean and unit variance



S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015

Batch normalization

- Normalize by channel-wise mean μ_c and standard deviation σ_c

$$\mathbf{Z} \in \mathbb{R}^{B\times W\times H\times C}$$

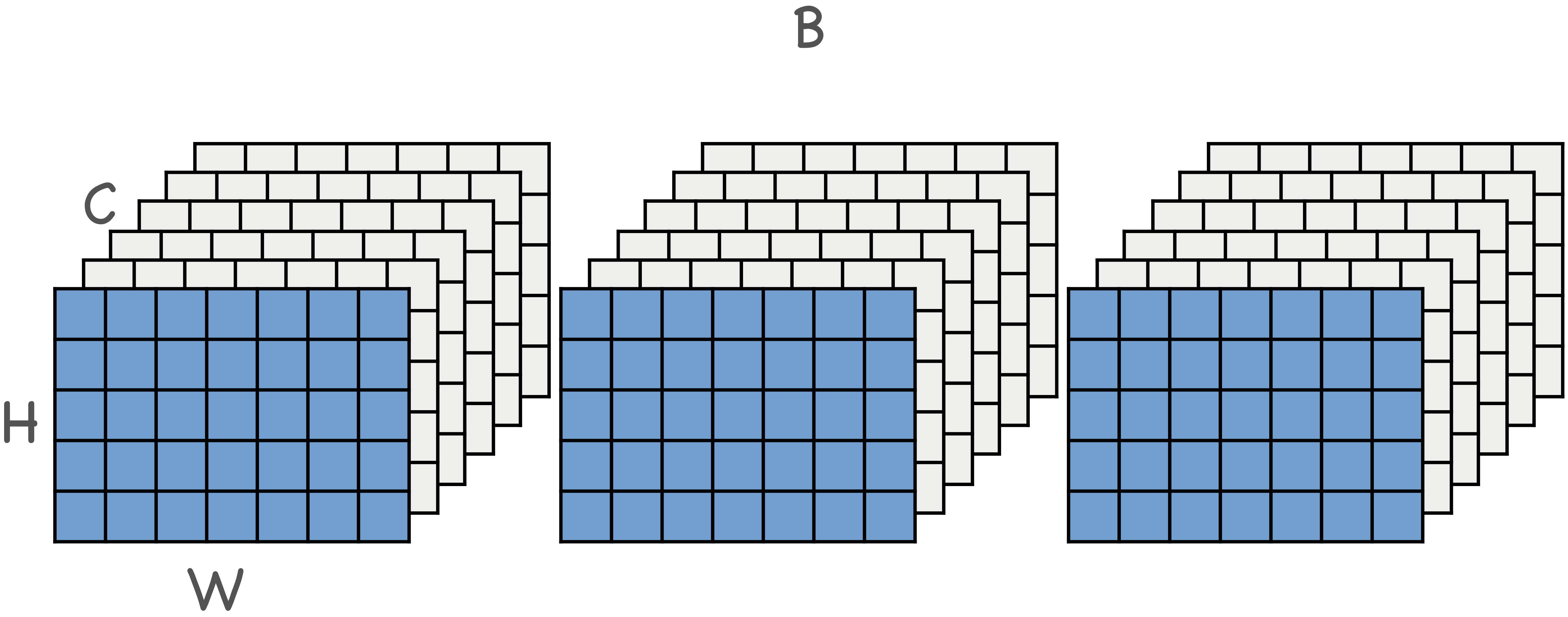
$$\downarrow$$

$$\boxed{\frac{\mathbf{Z}_{k,x,y,c} - \mu_c}{\sigma_c}}$$

$$\mu_c = \frac{1}{BWH}\sum_{k,x,y}\mathbf{Z}_{k,x,y,c}$$

$$\sigma_c^2 = \frac{1}{BWH}\sum_{k,x,y}(\mathbf{Z}_{k,x,y,c} - \mu_c)^2$$

What does batch normalization do?



What does batch normalization do?

- The good:
  - Regularizes the network
  - Handles badly scaled weights
- The bad:
  - Mixes gradient information between samples

$$\mathbf{Z} \in \mathbb{R}^{B\times W\times H\times C}$$

$$\downarrow$$

$$\boxed{\frac{\mathbf{Z}_{k,x,y,c} - \mu_c}{\sigma_c}}$$

$$\mu_c = \frac{1}{BWH}\sum_{k,x,y}\mathbf{Z}_{k,x,y,c}$$

$$\sigma_c^2 = \frac{1}{BWH}\sum_{k,x,y}(\mathbf{Z}_{k,x,y,c} - \mu_c)^2$$

Batch norm and batch size

- Large batch sizes work better
  - More stable mean and standard deviation estimates

$$\mathbf{Z} \in \mathbb{R}^{B\times W\times H\times C}$$

$$\downarrow$$

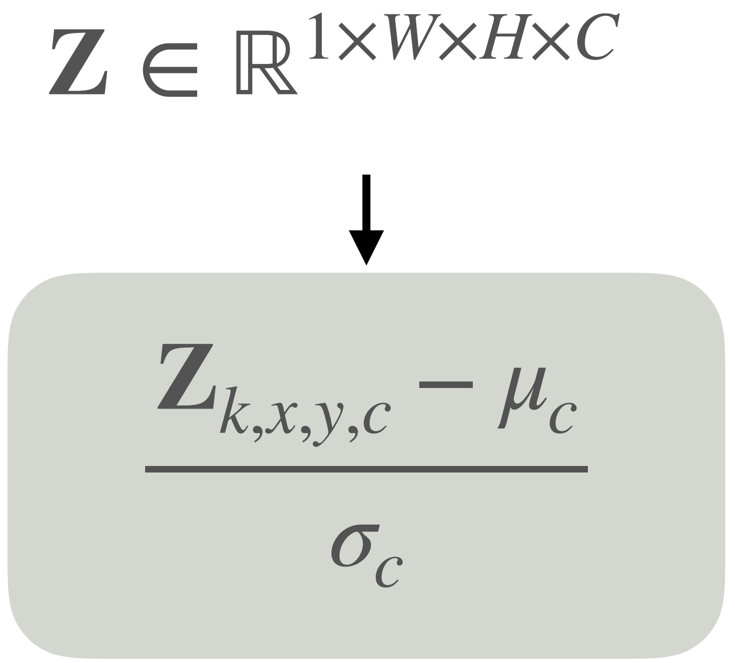
$$\boxed{\frac{\mathbf{Z}_{k,x,y,c} - \mu_c}{\sigma_c}}$$

$$\mu_c = \frac{1}{BWH}\sum_{k,x,y}\mathbf{Z}_{k,x,y,c}$$

$$\sigma_c^2 = \frac{1}{BWH}\sum_{k,x,y}(\mathbf{Z}_{k,x,y,c} - \mu_c)^2$$

Batch norm at test time

- Compute mean and standard deviation on training set using running average



$\mathbf{Z} \in \mathbb{R}^{1\times W\times H\times C}$

$$\frac{\mathbf{Z}_{k,x,y,c} - \mu_c}{\sigma_c}$$